CA Condensates의 SDI 서어비스를 위한 프로파일 作成法

<韓國科學技術情報센터 提供>

1. 序　論

急增하는 國內 情報需要에 效果的으로 對處하기 위하여 韓國科學技術情報센터에서는 化學分野의 著名한 旣存 데이터 베이스인 CA Condensates(CAC)와 그 檢索프로그램을 導入하여 컴퓨터에 의한 最新情報檢索서어비스(SDI)를 75年 7月 부터 本格的으로 開始하게 되었다.

이 서어비스를 利用하면 연구개발 활동에 필요한 最新技術情報를 신속정확하게 網羅的으로 간편하게 入手할 수 있어 先進諸國에서도 이 서어비스의 利用者가 급격히 증가하고 있다.

韓國科學技術情報센터에서는 앞으로 데이터 베이스를 확장하여 化學分野뿐만 아니라 科學技術 全分野에 대하여 이 서어비스를 實施할 計劃이다. 이 서어비스를 利用하려면 利用者마다 자기가 願하는 主題을 컴퓨터가 취급할 수 있는 形態(profile)로 만들어서 등록해 두어야 한다.

따라서 本稿에서는 利用者가 CAC 데이터 베이스를 理解하는 것은 물론 프로파일의 作成方法(profilling)을 習得할 수 있도록 마련하였다.

그러나 프로파일 作成에 필요한 모든 知識을 다 記述한다는 것은 그다지 쉬운 일이 아니다. 왜냐하면 紙面관계도 있겠지만 특히 각개인 情報要求에 따라 變化하는 因子의 다양성때문이다. 그러나 本稿에서는 그동안 經驗을 토대로 하여 利用者들로 하여금 적은 노력으로 효율적인 프로파일을 作成할 수 있도록 하였다. 되도록 많은 利用者들이 이를 익혀서 硏究開發活動에 큰 도움이 되기를 바란다

2. CAC 데이터 베이스

1) 範 圍

CA(Chemical Abstracts)는 世界 各國에서 刊行하고 있는 化學 化學工學 및 이와 관련있는 모든 分野의 學術雜誌 特許 技術報告書 政府刊行物 單行本 등의 抄錄이 연간 약 40만건이나 收錄되는 世界最大의 抄錄誌로 週 1回 册子形으로 出版되고 있다. CAC는 CA와 같이 每週 磁氣테이프 形態로 作成되는 것이며 收錄事項도 同一하나 다만 抄錄대신 일련의 重要語(Keyword)로 바꾸어 놓은 것만이 다르다.

2) 데이터 베이스 코오드

CAC 프로파일에는 다음과 같은 코오드가 쓰인다.

C_1 홀수週 발행분(有機化學／生化學)

C_2 짝수週 발행분(高分子 物理 分析과 應用化學 化學工學)

C_1 C_2 홀수와 짝수週 발행분

3) 檢索分野의 選擇

CAC의 檢索領域과 프로파일 검색 항목 카아드에 使用할 檢索領域 코오드는 다음과 같다.

領 域	檢索코오드
著 者	A
作業場所	W^+
標 題	H

發行日	D^+
發行物 分類코오드	P^+
CODEN	C^+
CA細項／細目	X
言 語	L^*
重要語	K^-
標題 및 重要語	T

$^+$표를 한 코오드는 使用하는 일이 극히 드물다.

4) 檢索領域의 構造

(1) 著 者

한 개의 레코드에 10名까지의 著者와 共著者를 포함시킬 수 있다.

人名은 姓을 먼저 쓰고 이름은 첫자만 따서 나중에 쓴다. 프로파일은 著者檢索에 있어서 너무 세밀하지 않게 하는 것이 좋다.

例를 들면 SMITH D*라고 하는 것이 가장 안전한 方法이다(*표는 用語줄임을 나타냄).

(2) 文獻標題

문헌표제는 英語가 아닌 경우 英語로 번역하여 편집한다. 이領域과 重要語領域에서는 美國式 철자법을 사용한다(例 Sulfur). 略號化學元素記號 및 化學式은 一般的으로 全 名稱으로 확장 사용한다.

(3) CA 細項／細目

CAC에 포함된 문헌은 CA와 마찬가지로 80개 細項(Section)으로 分類되어 있다. 各 문헌은 이들 중 어느 細項 하나에 들어가 있으며 이를 더 細分하여 細目에 分類되어 있다.

細項과 細目은 다음과 같이 수록되어 있다. CA mmmnnn 여기서 mmm은 細項을 세자리로 보아 001—080 중의 하나이고 nnn은 細目인데 각 細項別로 그 內容이 다르다.

例를들면 合成有機藥品에 관련된 문헌은 細項 64와 細目 3이기 때문에 CA 064003으로 表示되어 있다.

*L파라미터는 검색을 특정 細項으로 제한할때 써야하지만 細項／細目領域은 검색코오드 X를 쓴 표준 파라미터를 쓸 수 있다.

例를들면 合成有機藥品에 관련되는 記事를 찾아내기 위한 아주 좁은 검색인 경우 다음과 같이 쓴다.

(A) $\begin{matrix}*L\\ *P\ X\end{matrix}$ C2 $\begin{matrix}64\\ CA\ 064003\end{matrix}$

또는 간단히 (B) *P X CA064003이라고 쓰면 된다.

그러나 (A)가 바람직하다. 다른 방법으로도 찾을 길이 많지만 가능한 한L파라미터를 使用하여 찾는 길이 확실하다고 할 수 있다.

(4) 重要語

이 중요어의 영역에는 숫자적으로 制限은 없다. 重要語는 한개의 단어일 수도 있고 두 개 이상의 單語일 수도 있다. 이 重要語들은 標題 抄錄 및 原文에서 선택된 것도 있고 추가하는 수도 있다. 重要語에 쓰이는 各種略號는 美國化學會가 認定한 것에 限한다. 이러한 낱말은 디소오러스(Thesaurus)나 사전에서 선택한 것이 아니기 때문에 利用者는 스스로의 이용경험을 통하여 검색항목을 판단하여야 한다.

(5) 標題와 重要語

T檢索이란 검색항목으로 표제와 중요어語에서 찾도록되어 있다.

3. 프로파일의 구조

CAC 프로파일의 例를 表 1에서 紹介하였다. 프로파일을 作成하기 전에 利用者는 프로파일의 구조와 검색항목 사이에 적용되는 論理에 대해서 充分히 알아야 한다.

1) 制御카아드와 檢索項目카아드

프로파일은 7種의 制御카아드(*R,*N,*D,*M,*T,*E,*Z)와 2種의 검색항목카아드(*L,*P)로 되어 있는데 모든 制御카아드는 반드시 순서있게 나열해야 한다.

(1) *R카아드

欄 1—2 *R
3—4 版 數
5—12 利用者코오드
13—22 데이터 베이스의 종류
23 프로파일의 形式
24 出力部數
25 프로파일 상태
41 프로파일 작성자의 略字

(2) *N카아드(3장까지 쓸 수 있음)

欄 1—2 *N(첫장에 한함)
5—76 利用者의 住所 姓名

(3) *D카아드(3장까지 쓸 수 있음)

欄 1—2 *D(첫장에 한함)

5—76 요구주제에 관한 說明

(4) *M카아드

欄 1—2 *M

5—8 검색문헌의 最大프린트 부수

(5) *T카아드

欄 1—2 *T

5—8 基準加重値

(6) *E카아드

欄 1—2 *E(프로파일의 끝을 나타냄)

(7) *Z카아드

欄 1—2 *Z

(8) *L카아드

欄 1—2 *L

7—8 데이터 베이스의 코오드

12—14 一般的으로 空欄

16—75 데이터 베이스에서 細分한 項目의 코오드

*L카아드를 한장 이상 쓸 수 있다.

*L표시는 첫장에만 쓰되 각 카아드의 7—8欄에 데이터 베이스 코오드는 반드시 記入하여야 한다.

(9) *P 카아드

欄 1—2 *P

4 검색분야 코오드

5—8 純論理 프로파일인 경우엔 빈칸 加重 프로파일인 경우 항목의 加重値

12—14 NOT 論理項目인 경우 NOT, 그밖의 경우 빈칸

16—75 검색항목

21—75 IGNORE 論理項目

2) 프로파일 論理

表 1 純論理 프로파일

```
0000 *ROICNAPHI C1C2 11P GLJ
0010 *N MRBB BLACK
0020     CSIRO DIVISION OF ANIMAL PHYSIO
         LOGY
0030     BLACKTOWN N.S.M
0040 *D INFORMATION RELATING TO THE
         LOSS OF HAIR, WOOL OR
```

다음 4가지 論理가 적용된다.

(1) OR論理

파라미터內의 各 검색항목 사이에 적용된다.

項目의 어느 하나라도 해당되면 파라미터는 만족하게 된다.

(2) AND 論理

파라미터間에 적용된다. 全파라미터가 만족되어야만 프로파일이 만족되어 검색된다.

(3) NOT 論理

검색항목 카아드의 12—14欄에 制限되어 있다. 어떤 파라미터 內의 어떤 항목에 NOT論理가 附與되어 있다면 이 파라미터 內의 全項目이 컴퓨터에 의하여 NOT 論理가 부여된다. NOT論理 項目이 해당된다면 레코드는 즉각 無視된다.

(4) IGNORE 論理

프로파일의 IGNORE論理 구조는 다음과 같다.

CHROM IGNORE CHROMATOGRAPHY NOT 論理보다 약한 것이며 IGNORE 論理는 CHROMATOGRAPHY와 관계없이 된다. 그러나 계속해서 레코드를 檢討한다.

그래서 CHROMATES의 CHROMATOGRAPHY에 관한 문헌은 重要語群에서 CHROMATES란 項目이 있는 경우에만 검색된다. IGNORE 아래에 오는 項目은 관련이 있는 검색항목을 포함하여야 한다. 例를 들면 GENE IGNORE GEN라고 하면 된다.

4. 프로파일의 作成

1. 作成方法의 種類

프로파일을 作成하는 方法에는 다음의 2가지가 있다.

```
0050     FKEECE AR A RESULT OF ANY CHE-
         MICAL, METABOLIC,
0060     NUTRITIONAL OR ENVIRONMENTAL
         INFLUENCE
0070 *M 200
0080 *T   0
0090 *L  C1  1 2 3 4 5 6 7 8 9 13 14 15 17 18 29
             30 31 32 33 34
0100     C2  41 62 63
0110 *P  T   *HAIR*
0120             IGNORE
0130             ROOT HAIR*
0140         *WOOL*
```

```
0150          *FLEEC*
0160          *KERATIN*
0170          FUR
0180          *MITOTIC*
0190          DEPILAT*
0200 *P  T    DEPILAT*
0210          COSMETIC*
0220          REVIEW
0230          *NUTRI*
0240          *FOLLIC*
0250          HORMON*
0260          METAB*
0270          ROOT*
0280               IGNORE
0290               ROOT HAIR
0300          COPPER
0310          *PLOIDY
0320          SHEARING
0330          ALOPECIA
0340          DISULFID*
0350          AMINO ACID*
0360          RADIATION
0370          ZINC
0380          CHELAT*
0390          ENZYM*
0400          *PROTEIN*
0410          SULFHYDRYL*
0420          CHEMOTHERAP*
0430          *MOULT*
0440          DERMAT*
0450 *E
0460 *Z
```

(1) 純論理 프로파일(pure Logic profile)

純論理 프로파일의 例를 表 1에서 紹介하였다. 項目이 每 파라미터마다 매치된다면 純論理프로파일은 만족된다. 파라미터 內에서의 探索은 매치가 됨으로써 끝나고 컴퓨터는 다음 파라미터로 옮긴다. 그리고 全 파라미터가 만족되는 경우 레코오드가 검색된다. 純論理 프로파일은 一般的으로 加重 프로파일보다 빠르게 검색된다.

명확히 定義된 狹少한 관심사의 검색에 적절하다.

(2) 加重 프로파일(Weighted profile)

加重 프로파일의 구조는 純論理 프로파일과 같으나 다른것은 每 探索項目마다 同 項目카아드의 5—8欄에 加重値 플러스, 零 또는 (마이너스)를 준 것이다. 매치 과정에서 全項目에 대한 加重値를 合算한다. 純論理 프로파일의 경우와 같이 매치되기 前까지는 파라미터의 探索이 끝나지 않으므로 留意할 필요가 있다. 加重 프로파일의 全 파라미터가 만족된다면 즉 每파라미터에서 적어도 하나의 매치가 있다면 加重値를 合算하고 미리 設定해 둔 기준식과 같거나 이보다 크면 검색하게 된다. 加重檢索에서는 探索項目에 加重値를 −250으로 定하면 NOT 論理와 같은 結果가 된다. 이 項目으로서 매치되는 경우의 문헌은 무시된다. 다른 論理는 모두 定해진대로 적용된다. 加重探索에서 出力되는 문헌은 검색 加重値가 높은 순서대로 프린트 되는데 純論理 探索에서는 문헌번호의 순서대로 프린트 된다.

表 2. 加重 프로파일

```
10 *RO1CQAH2    C1        I1P        GLJ
20 *N DR G G GREEN
30      CSIRO DIVISION OF ANIMAL HEALTH
40      INDOOROOPILLY QLD
50 *D PATHOPHYSIOLOGY OF BABESIA AR-
      GENTINA AND B. BIGEMINA
60      INFECTIONS OF CATTLE PARTICULA-
      RLY PHARMACOLOGICAL
70      ASPECTS OF THE DISEASE AND THE
      VASCULAR SYSTEM
80 *M     200
90 *T     110
100 *P T  80        BABES*
110       90        KININ*
120       80        BRADYKININ*
130       90        KALLI*
140       80        SEROTONTI* TONI*
150       80        HISTMAIN*
160       50        INFLAMMAT*
170 *P T  50        PATHO*
180      -250       ANTAGON*
190    K  30        *PEPTID*
200       30        ANGIOTENSIN
210       20        REVIEW
220       30        SHOCK*
230       20        PYRETIC
240       30        BLOOD
250       30        EXUDAT*
```

```
260        30      VASO*
270        20      *GLOBULIN*
280        20      ANAPHYL*
290        30      VASCULAR*
300        30      TRYPTAMIN*
310        30      ASSAY
320        30      PROTOZOA*
330        30      TRYPANOSOMA
340        30      PLASMODIUM*
350         0      *KININ*
360         0      *HISTAMIN*
370        80      BABES*
380        30      BOVINE
390       -30      DRUG*
400 *E
410 *Z
```

2. 검색효율을 높이는 방법

프로파일을 作成할 때에 다음의 몇가지 方法을 使用하면 檢索效率을 높일 수 있다.

(1) 項目의 切斷(Term Truncation)

項目을 절단해서 검색하는 방법을 터엄 트런캐이션이라고 하는데 단어의 語幹만 가지고 찾으면 검색능력이 향상된다. 잘려진 검색항목에는 표를 하는데 語頭나 語尾에 붙인다. 기계로 처리하는 과정에서 單語 절단 표시만 하여 주면 컴퓨터는 *표한 자리에 다른 文字群이 와 있었던 것으로 取扱해 준다. 項目이 잘려져 있지 않으면 項目의 뒤에 빈칸이나 點이 올때까지 처리한다. 例를 들면,

```
Polymer
*Polymer
Polymer*
Polymer*
```

라고 쓰면

```
Polymer
Polymer, Copolymer 등
Polymer, Polymers 등
Polymer, Copolymerization 등
```

이 검색된다.

그러나 단어의 길이가 짧아진만큼 不適合한 문헌이 검색될 위험성이 커진다. 例를 들면 *ASE와 *OSE 酵素나 설탕 등의 語尾와 같다고 해서 이들로서 효소나 설탕에 관한 문헌을 검색할 수는 없다. 왜냐하면 Base, phase, Hose와 같은 일반적인 낱말들이 검색되어 버리기 때문이다. 또 Carbon * Oxide와 같이 중간에 *표를 사용할 수 없다.

(2) 필터 파라미터(Filter parameter)

探索效率은 프로파일에서 첫번째 파라미터의 검색항목의 數가 적을수록 더욱 최적화되는 것이다. 일반적인 규칙으로서 첫째 파라미터는 10個의 探索項目 以內로 제한하여야 한다. 그러나 때때로 不可能한 때가 있다. 이때 첫째 파라미터를 시험하기 전에 필터파라미터를 사용하여 出力에는 역효과가 없이 프로파일 효율이 向上된다. 例

```
*P T        NITROGEN
            NITROGEN FIX*
            NITROGENOUS
            AMMONI*
            NITRAT*
            NITRIT*
            NITRIF*
                IGNORE
                *NITRIFICAN*
            15N*
```

이것은 목장에서의 질소 경제에 관한 프로파일의 첫째 파라미터이다. 이런 경우에 아래와 같은 필터 파라미터를 挿入함으로서 出力에 逆效果가 없이 프로파일 효율이 向上된다.

```
*P T        NITR
            AMMONI
            15N
```

이 필터 파라미터에서는 3개의 探索項目만으로 부적합한 레코오드를 일단 除去하고 餘分의 레코오드는 위의 파라미터로서 대조하면서 제거해 나간다. 이런 形式의 필터 파라미터를 도입하면 探索時間을 40%까지 절약할 수 있다.

(3) *L파라미터의 使用

이미 說明한 바와 같이 CAC는 다음과 같이 80個 細項으로 分類되어 있다.

홀수週(C1) 細項 1—34

짝수週(C2) 細項 35—80

*L을 사용하면 특정 細項만 찾기 때문에 시간 절약이 가능하다.

*L C1 17 20 26 31

라고 쓰면 C1에 있는 34개 細目에서 4개 細項만 찾는 것이 되며

L* C1 NOT 33

라고 쓰면 細項 1—34까지에서 33만 除外하고 모두 찾는다는 뜻이 된다.

(4) 파라미터의 數와 順位

프로파일에서 使用되는 파라미터의 數와 그 順位는 效率에 미치는 영향이 크다. 파라미터의 수가 많을수록 요구주제를 具體的으로 정확하게 찾을수는 있겠지만 너무 細分化되어서 出力이 적어지고 때로는 프로파일 作成者는 되도록 적은 파라미터를 使用하여 要求主題를 正確하게 찾도록 하는것이 최선의 방법이다. 또한 2개 이상의 파라미터를 사용할 경우 어느 파라미터를 먼저 사용할 것인가 하는 문제도 중요하다 이 경우엔 使用빈도수가 적은 探索項目이나 그의 파라미터를 먼저 사용하는 것이 原則이다.

5. 프로파일의 評價와 修正

1) 評　　價

프로파일의 檢索結果는 다음과 같은 事項을 檢討하여 評價할 수 있다.

(1) 手作業으로 檢索된 문헌중에서 검색되지 않은 文獻의 數.

(2) 手作業으로 검색되지 않았던 문헌 중에서 檢索된 適合文獻의 數

(3) 檢索된 不適合 文獻의 數

探索범위의 깊이는 이 단계에서 決定되어야 하며 이것은 요구정보의 用途에 따라 左右된다. 즉 利用者가 特定分野의 연구에 從事한다면 그는 部分的으로만 적합하다고 할지라도 그 分野에 관하여 出版된 것이라면 모든 문헌을 요구할 것이다. 그러나 높은 再現率을 요구하는 探索은 많은 不適合 文獻을 검색한다는 것을 잊지 말아야 한다. 그러므로 꼭 필요한 주제만을 요구할 때에는 높은 精度의 探索을 함으로써 不適合한 문헌의 검색을 줄일 수 있다.

2) 修　　正

프로파일을 修正하는 경우는 檢索結果의 상태에 따라 다음과 같이 4가지로 나눌 수 있다.

(1) 檢索된 문헌이 없을 경우 : 檢索項目의 범위를 넓혀 프로파일을 再構成한다.

(2) 검색된 문헌의 數가 手作業에 의한 검색문헌의 數보다 너무 적은 경우 : 手作業에 의해 抽出된 검색항목의 論理的 組合에 유의하여 프로파일을 재구성한다.

(3) 약간의 不適合한 문헌과 함께 手作業에 의해 검색 되었던 문헌이 검색되는 경우 : 修正할 필요없이 실제적으로 사용한 프로파일의 形態로 바꾼다.

(4) 手作業에 의해 검색되었던 문헌이 많은 부적합한 문헌과 함께 검색되는 경우 : 不適合한 문헌을 검토하여 서로 어떤 共通되는 現象을 갖고 있는가를 考察하여 불필요한 문헌을 제외시킬 수 있도록 프로파일을 재작성한다.

이러한 과정은 效果的인 프로파일을 作成할 때까지 反復한다. 끝맺으면서 한국과학기술정보센터에서 처음 시작하는 CAC SDI 서어비스는 科學技術情報利用者들에게 놀랄만한 喜消息이다. 爆增하는 정보 속에서 자기가 필요한 최신기술정보를 신속정확하게 網羅的으로 간편하게 入手할 수 있기 때문이다.